Big Data Engineering Term Project

Phase 1: Star Schema

—

Data Warehouse for DVD Rental Business

Database Source Name: Sakila Database
Business Area: Sales and Business Development

Submitted By:

- **Mahmoud Elshikha** (22-101193)
- **Yousef Walid** (22-101048)
- **Mohamed Mostafa** (22-101203)
- **Mohamed Ibrahim** (22-101058)

Instructors: Prof. Hoda & Eng. Maha Sayed
Course: Big Data Engineering
Date: 3/12/2025

Business Goals

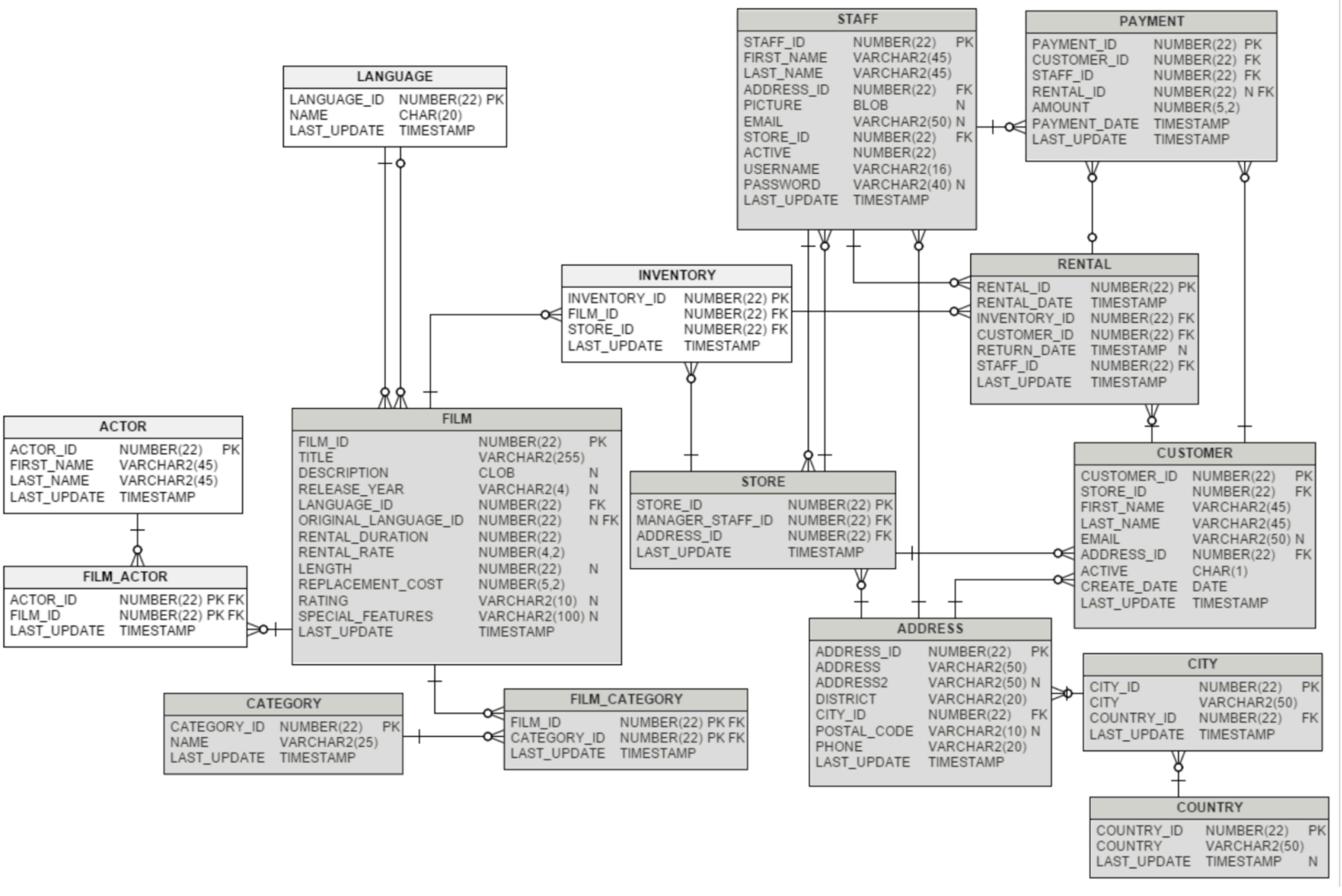
This data warehouse project focuses on **sales performance analysis** of a DVD rental business:

1. **Boost Revenue Through Sales Insights**
    - Analyze total revenue generated from rentals, late fees, and payments across time.
    - Compare film categories and individual titles to determine high-performing genres.
    - Understand store-wise and region-wise sales performance.
2. **Identify High-Value Customers**
    - Track customer purchase behavior based on payment amounts and frequency.
    - Segment customers into tiers based on spending to offer targeted promotions.
3. **Uncover Seasonal Trends in Sales**
    - Use date-based insights to analyze monthly, quarterly, and yearly sales trends.
    - Prepare for peak seasons and allocate resources accordingly.

ERD

The following tables from the Sakila database are relevant to our analysis:

1. **Rental:** The fact table that records all rental transactions. Essential for analyzing rental frequency, trends, and customer behavior.
2. **Payment:** Provides financial data, such as total payments and late fees.
3. **Customer:** Contains customer details such as name, email, and active status.
4. **Film:** Needed for movie-specific analytics like popularity, rental rates, and duration.
    a. **Film_Category**
    b. **Category**
5. **Category:** Helps in genre-based analysis (which movie genres perform best?).
6. **Store:** Allows location-based analysis, identifying which stores perform best.
7. **Address:** Provides location details like city and country (linked to the store table).
    a. **City**
    b. **Country**
8. **Staff:** Used to get the manager's staff ID.



Star Schema (Dimensional Model)

Fact Table: Fact_Rental

Column Name	Description
customer_id (FK)	Links to the Customer dimension.
film_id (FK)	Links to the Film dimension.
store_id (FK)	Links to the Store dimension.
staff_id (FK)	Links to the Staff dimension.
date_id (FK)	Links to the Date dimension.
rental_duration	Calculated as return_date - rental_date in ETL.
payment_amount	Total payment for the rental.
late_fee	Calculated during ETL. Charge 1.00 base fee if it's late. Then, for each additional late day, apply a 5% increase to the base. ● So 1 day late → `1.00 + 1.00 * 0.05 = 1.05` ● n days late → `1 + 0.05 * n`

Dimension Tables: Dim_Customer

Column Name	Description
customer_id (PK)	Unique customer identifier.
first_name	First name of the customer.
last_name	Last name of the customer.
email	Customer's email address.
City	City of customer
Country	Country of customer
active_status	Indicates if the customer is active.

Dimension Tables: Dim_Film

Column Name	Description
film_id (PK)	Unique film identifier.
title	Movie title.
release_year	Year the film was released.
rental_rate	Price to rent the movie.
length	Duration of the movie in minutes.
category_name	Genre name (e.g., Action, Comedy).

Dimension Table: Dim_Store

Column Name	Description
store_id (PK)	Unique store identifier.
manager_staff_id	Staff ID of the store manager.
city	Name of the city.
country	Country name.

Dimension Table: Dim_Staff

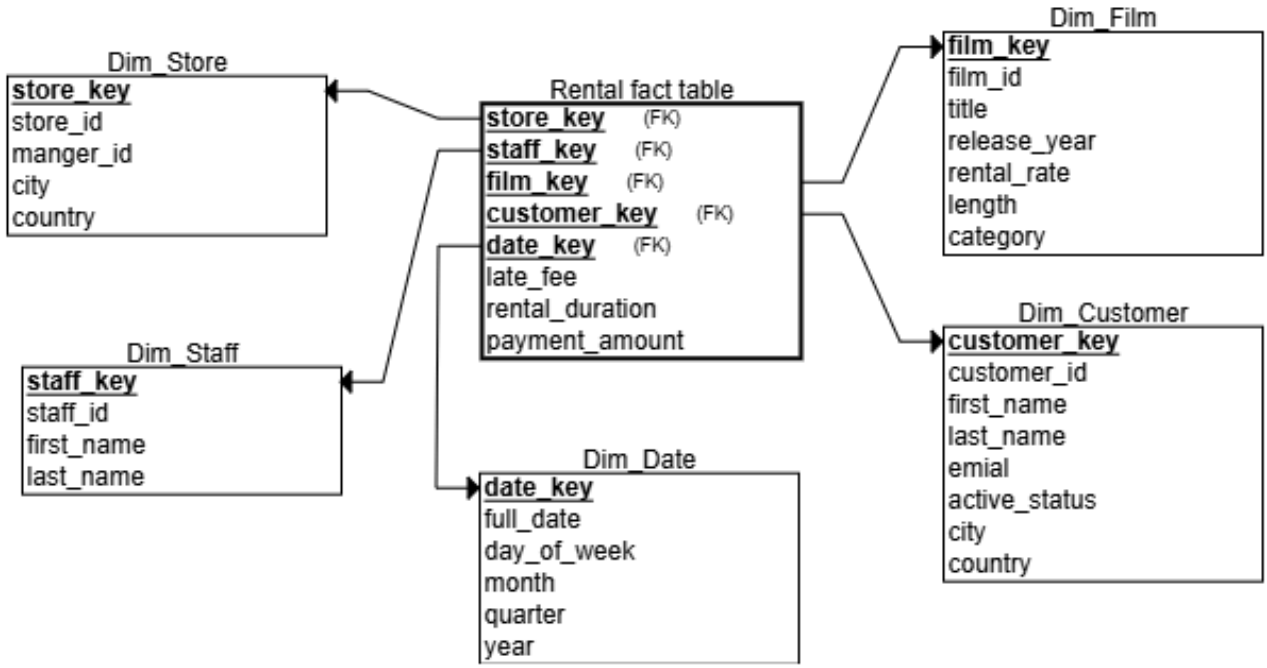
Column Name	Description
staff_id (PK)	Unique staff identifier.
first_name	First name of the staff member.
last_name	Last name of the staff member.

Dimension Table: Dim_Date

Column_Name	Description
date_id (PK)	Unique address identifier.
full_date	11-11-2003 format

day_of_week	7 days of the week
month	Which month
quarter	First-second-third-fourth
year	Which year

Schema Diagram



Schema Description (Dimensions, Dimension Levels, and Measures)

Dimension Tables and Levels

1. **Dim_Customer** (Stores customer details)
- **Levels:** Customer → Location
- **Attributes:** customer_id (PK), first_name, last_name, email, active_status

2. **Dim_Film** (Stores film details)
- **Levels:** Film → Genre
- **Attributes:** film_id (PK), title, release_year, rental_rate, length, category_name

3. **Dim_Store** (Stores store details)
- **Levels:** Store → City → Country
- **Attributes:** store_id (PK), store_name, manager_staff_id, city, country

4. **Dim_Staff** (Stores staff details)
- **Levels:** Staff
- **Attributes:** staff_id (PK), first_name, last_name

5. **Dim_Date** (Stores time-related information)
- **Levels:** Date → Month → Quarter → Year
- **Attributes:** date_id (PK), full_date, day_of_week, month, quarter, year

Measures (Quantitative Data in the Fact Table)

rental_duration: Measures how long a film was rented before being returned.
payment_amount: Total revenue generated per rental.
late_fee: Additional revenue from overdue rentals.